A MIXED-SOURCE MODEL FOR SPEECH COMPRESSION AND SYNTHESIS

J. Makhoul, R. Viswanathan, R. Schwartz and A.W.F. Huggins

Bolt Beranek and Newman Inc.
Cambridge, Mass. 02138

ABSTRACT

This paper presents an excitation source model for speech compression and synthesis, which allows for a degree of voicing by mixing voiced (pulse) and unvoiced (noise) excitations in a frequency-selective manner. The mix is achieved by dividing the speech spectrum into two regions, with the pulse source exciting the low-frequency region and the noise source exciting the high-frequency region. A parameter F_c determines the degree of voicing by specifying the cut-off frequency between the voiced and unvoiced regions. For speech compression applications, F_c can be extracted automatically from the speech spectrum and transmitted. Experiments using the new model indicate its power in synthesizing natural sounding voiced fricatives, and in largely eliminating the "buzzy" quality of vocoded speech. A functional definition of buzziness and naturalness is given in terms of the model.

1. INTRODUCTION

Perhaps the single most important decision to be made in a pitch-excited speech compression system (vocoder) is the voiced/unvoiced (V/U) decision. Errors in this decision are readily perceived by the ear as a degradation of speech quality, and may also be accompanied by a loss in intelligibility. Yet, even if the V/U decision were somehow to be made "perfectly", the synthetic speech would continue to exhibit a distinct lack of naturalness, exemplified by a certain "buzziness" and a "lack of fullness." These characteristics are symptoms of the inadequacy of the binary V/U excitation model.

This paper explores the excitation problem in speech synthesis and presents a simple mixed-source model that allows for a degree of voicing. The new model is capable of producing more natural sounding speech; it seems to largely eliminate the buzziness problem and recover much of the fullness in the speech. In addition, it promises to reduce the adverse effects of voicing errors. A review of previous research relating to this model is given in a later section.

2. BASIC SYNTHESIS MODEL AND TERMINOLOGY

Throughout this paper, we shall assume the basic synthesis model shown in Fig. 1. In this model, a time-varying excitation signal excites a time-varying spectral shaping filter, the output of which is the synthetic speech. The excitation signal is assumed to have a flat spectrum, so that the spectral envelope of the synthetic speech is determined completely by the spectral shaping filter. Furthermore, we shall assume this model to hold for any type of synthesis, whether as part of a vocoder system or a synthesis system. In fact, we wish to argue below that our proposed source model is indeed adequate for both applications.

Restricting the excitation to have a flat spectrum necessarily limits us to two types of excitation: deterministic (pulse) or random (noise).

a) Pulse Source (Buzz)

The deterministic excitation is, in general, the impulse response of an all-pass filter, which we shall call an all-pass signal or pulse. The most trivial form of an all-pass pulse is a single impulse. When the pulse source produces a sequence of pulses separated by a pitch period, it is known as a buzz source. (Note that a single pulse could be used in the synthesis of the burst in a plosive sound [1]. However, the burst can also be synthesized using the noise source. We shall assume the latter in this paper; the pulse source will be used exclusively for buzz excitation.)

b) Noise Source (Hiss)

The random noise excitation may be the output of a random number generator. Generators with either a uniform or Gaussian probability distribution are readily available and are quite adequate. The noise source is also known as a hiss source.

Whether the actual excitation is buzz or hiss, or a combination of the two, one must always make sure that the excitation has a flat spectrum. We shall now describe how one might derive an appropriate source model by inspecting short-time speech spectra.

3. THE "IDEAL" SOURCE

For some particular speech signal, one can remove the short-time spectral envelope by appropriate inverse filtering, as shown in Fig. 2. The inverse filter A(z) can be obtained by cepstral techniques [2] or through the use of linear prediction [3]. The residual signal e(t) will then have a nominally flat spectrum. If in Fig. 1, the excitation u(t) is identical to the residual e(t), and the synthesis filter H(z) is the inverse of A(z), then the synthetic speech s'(t) will be identical to the original signal s(t).

However, for synthesis purposes, the synthetic signal need only sound like the original, and need not be identical to it. In addition, we need to manipulate the source pitch and to minimize the number of bits needed to represent the source. In order to accomplish this task, we first make use of an important property of speech perception, namely

CH1285-6/78/0000-0163$00.75©1978IEEE

that it is relatively insensitive to the short-time phase. Therefore, in order to model the residual e(t) to meet our requirements, we need only look at its spectrum and, except for pitch, disregard its phase structure for the moment.

Fig. 3 shows the signal power spectrum of 25.6 ms of a 10 kHz sampled signal in the middle of the vowel [I] in the word "list", and the corresponding residual spectrum. The residual was obtained by inverse filtering the speech signal with a 20th order linear prediction inverse filter. If somehow one could generate an excitation u(t) whose spectrum is identical to the residual spectrum, the synthetic speech would then sound (almost) the same as the original.

Therefore, our aim in developing source models will be to obtain an excitation spectrum that is as close as possible to the residual spectrum. Furthermore, we wish to obtain such an excitation spectrum using only the buzz and hiss sources described in Section 2. The source models will stem naturally from examining the characteristics of residual spectra.

4. CHARACTERISTICS OF RESIDUAL SPECTRA

In Fig. 3, the residual spectrum shows a clear periodicity up to about 3.5 kHz, and a lack of periodicity above that frequency. The periodicity corresponds to harmonics of the pitch fundamental frequency. By looking at residual spectra of other sounds it becomes amply clear that the existence of aperiodic frequency bands in sonorant sounds is quite common. While in Fig. 3 one can identify only two bands, it is possible to have several periodic and aperiodic adjacent bands in 5 kHz. For more examples, the reader is referred to the work of Fujimura [4], who studied voice aperiodicity by examining short-time signal spectra.

Partial devoicing of certain sounds is well-known from physical considerations. For example, the devoicing of [z] above about 1 kHz is well recognized and has long been taken advantage of in the synthesis of more natural voiced fricatives. On the other hand, it is also known that in the production of the tense front vowel [i], the constriction may become narrow enough to generate some turbulence, which is seen as devoicing of frequencies above about 3 kHz. However, most synthesizers to date have not taken advantage of this fact.

In addition to the foregoing types of sources of devoicing, Fujimura [4] has hypothesized that some of the spectral devoicing may be due to aperiodicities or irregularities in the vocal-cord movement. We have noticed that spectral devoicing often occurs during transitions between different sounds, including sonorant-sonorant transitions. In contrast to the examples given in the previous paragraph, we believe that the spectral devoicing due to vocal-cord irregularities and/or spectral transitions, may in fact be an artifact of the spectral estimation process. Whether such devoiced regions should be synthesized using a noise source is questionable.

In conclusion, residual spectra may be completely periodic (voiced), completely aperiodic (unvoiced), or may contain regions that are periodic and others that are aperiodic. The question now is how to model such spectra using the buzz and hiss sources.

5. PROPOSED SOURCE MODEL

One reasonable source model would divide the spectrum into a number of bands. Each band would then be excited by the buzz source if the band is considered periodic, and by the hiss source if the band is considered aperiodic. Fujimura [4] used a 3-band model in his experiment, and reported an improvement in speech naturalness. However, given our observations that spectral aperiodicities may not necessarily result from turbulent excitations, we have chosen a different model. In our model, we shall consider all spectral aperiodic regions that are in between two periodic regions to be in fact periodic. In other words, only the band above the periodic region with the highest frequency will be considered to be aperiodic and generated by a turbulent source. Our reasons for this choice are twofold: (a) Turbulent sources are more likely to excite higher frequencies; and (b) Excessive devoicing can be as degrading to quality as excessive voicing.

The resulting model is shown in Fig. 4. It is a mixed-source model with the buzz source exciting a time-varying low-frequency region of the spectrum, and the hiss source exciting the remaining high-frequency region. The selective excitations are realized by passing the pulse excitation through a low-pass filter with cutoff F_c, and the noise excitation through a high-pass filter with the same cutoff frequency F_c. The outputs of the two filters are then added, multiplied by the source gain and applied to the spectral shaping filter as the excitation signal. The model, then, has only two parameters: the cutoff frequency F_c, and the pitch period τ when $F_c > 0$. Since small changes in F_c are not perceptible, it is sufficient to quantize F_c into 2-3 bits for transmission purposes.

6. IMPLEMENTATION

a) Extraction of Source Parameters

The only difference between parameter extraction for the new source model and traditional pitch extraction is that the V/U binary decision has been replaced by the determination of a multi-valued parameter F_c in our model. The extraction of the pitch period is unchanged. Pitch period determination is relatively straightforward; many schemes exist that are quite adequate.

Just as V/U decision algorithms have proliferated, many algorithms will be developed that attempt to compute F_c in a perceptually satisfactory manner. The method we have chosen thus far is a peak-picking algorithm on the signal spectrum. The algorithm determines periodic regions of the spectrum by examining the separation between consecutive peaks and determining whether the separations are the same, within some tolerance level. F_c is taken to be the highest frequency at which the spectrum is considered to be periodic.

b) Filter Implementations

In our initial implementation we rounded the value of F_c to the nearest 500 Hz. Therefore, we needed lowpass and highpass filters with cutoff frequencies separated by 500 Hz. The filter designs were then stored and used in the synthesis as the need arose.

For each value of F_c, the 3 dB points for the lowpass and highpass filters were designed to be equal to F_c, in order that the spectrum of the

final excitation may be as flat as possible. The roll-off of the filters was considered to be of secondary importance, but should not be very sharp in any case. We considered FIR (finite impulse response) as well as recursive (low order Butterworth) filters. Both types of filters gave similar perceptual results.

7. RESULTS

Using the implementation described in Section 6, we compared the resulting syntheses to those using the binary V/U model in the context of a linear prediction (LPC) vocoder. A number of sentences from male and female speakers [5] were used in comparing the two analysis-synthesis systems. No quantization of parameters (except for F_0) was performed. One of the sentences had a concentration of fricative sounds "His vicious father has seizures," and another was a nonnasal sonorant sentence "Why were you away a year, Roy?" Other sentences were more general. With the V/U source, the fricative sentence sounded particularly buzzy for both male and female speakers, while the sonorant sentence was judged as buzzy only for low-pitched male speakers. The buzziness in both sentences was greatly reduced when using the mixed-source model. In general, the buzziness was always reduced with the new model. However, for some sentences the new synthesis produced certain small background noises. Upon careful listening, it was determined that some of those noises were present in the V/U synthesis but were masked by the buzziness. The other noises may be due to inaccurate determination of F_0 and/or to the particular implementation of the model.

Overall, listeners thought that the new model performed better for female speakers (a pleasant surprise, for a change). The new synthesis was "raspier" and more in line with female speech which is considered to be more breathy than male speech.

A number of listeners reported that the new synthesis had a certain "fullness" that was absent with the V/U synthesis. We interpret this as an indication of the greater naturalness resulting from the new model.

8. REVIEW OF RELATED WORK

The only other work we know of where mixed excitation was used with LPC vocoders was that of Itakura and Saito [6]. But there, the two sources excited the whole spectrum simultaneously, with the "degree" of voicing being controlled by the relative <u>amplitudes</u> of the sources. The results were not encouraging [7].

After the development of our model over two years ago, we became aware of Fujimura's work [8,4], who as far as we know, was the first to suggest and test a frequency-selective mixed-source model. His work, which we mentioned earlier, was performed in the context of a pitch-excited channel vocoder. During the writing of this paper, Fujimura brought to our attention his other work with Kato et al. [9], where a variable cut-off frequency like ours was employed, but using a different algorithm to determine the cut-off. The work was done with a hybrid voice-excited and pitch-excited channel vocoder, and they reported excellent results. Coulter [10] used mixed excitation for the synthesis of voiced fricatives; however, the cut-off between the low and high frequency bands was fixed.

In speech synthesis, mixed excitation has been used routinely for the synthesis of voiced obstruents (see, for example, [1,11]). The parallel formant synthesizer of Holmes [1] allows for variable mixed excitation, and was especially used in transitions between unvoiced and voiced sounds. Upon careful reading, it became clear to us that the spirit of Holmes' synthesizer is similar to ours, except that the controls in his case are more complicated. A more recent hardware synthesizer by Strube [12] allows for mixed excitation using a single variable RC-circuit.

There have been numerous attempts at reducing buzziness by changing the shape of the pulse in voiced excitation, but to no avail. Recently, Sambur et al. [13] reported a reduction in buzziness by changing the pulse width to be proportional to the pitch period. Unfortunately, changing the pulse width changes the excitation spectrum; the effect is that of a variable lowpass filter. Spectrally flattening the pulse before excitation cancelled the reduction in buzziness [14].

9. DISCUSSION

a) Buzziness and Naturalness

It is interesting that the mixed-source model appears to reduce two seemingly different types of buzziness: the buzziness in voiced fricative synthesis, and the buzziness in sonorant synthesis associated mainly with low-pitched voices. Our hypothesis is that <u>the two types of buzziness</u>, in fact, <u>result from the same process</u>: <u>that of an excess in buzz source excitation</u>. Thus, our general rule is that:

too much buzz → "buzziness"
too much hiss → "breathiness" or "raspiness"

where the arrow is to be read as "results in". If more of the spectrum is excited by the buzz source than is necessary for naturalness, the result is buzziness. Similarly, if there is more hiss excitation than is necessary for naturalness, the result is breathiness or raspiness. This leads us to a functional definition of naturalness, as it relates to mixed excitation:

> Naturalness is achieved by that proper mix of buzz and hiss excitations that leads to a synthesis that is neither buzzy nor breathy or raspy.

b) Modulation and Naturalness

Certain synthesizers, such as that of Klatt [11], modulate the hiss source by the buzz source for the synthesis of voiced fricatives. While it is known that the noise source in the vocal tract is in fact modulated by the vocal cord output, it is not clear that such modulation is necessary for achieving naturalness in synthetic speech. Whatever effect modulation has, it appears to be of a secondary nature. The synthesizer of Holmes [1] does not contain any modulation, and he reported very natural speech synthesis. Although initially we included modulation in our model, it is our opinion at this point that source modulation is not necessary for natural synthesis, and therefore we have decided not to incorporate it as part of the model.

c) Phase and Naturalness

It is generally agreed that proper phase determination of buzz excitation should lead to more natural synthesis. Furthermore, such phase cannot be in the form of some "optimal" pitch pulse shape. The phase must change from one pitch pulse to the next in some appropriate manner. Thus far, our model calls for an all-pass pulse, but does not specify the phase. Exactly how the phase should change between pulses is a subject for future research.

10. CONCLUSION

We have presented a frequency-selective mixed-source excitation model for use in <u>both</u> speech compression and speech synthesis. The model has a single continuous parameter, F_c, which divides the spectrum into two regions, with the buzz source exciting the low frequency region below F_c, and the hiss source exciting the high frequency region above F_c. Naturalness (no buzziness or breathiness) is achieved by the proper mix of the two sources, i.e., by the proper determination of F_c.

ACKNOWLEDGMENTS

The authors wish to thank K.N. Stevens for many discussions, especially during the initial development of the model. One of the authors (JM) had a useful discussion with L. Rabiner concerning the different types of buzziness in vocoded speech. This work was sponsored by the Information Processing Techniques branch of the Advanced Research Projects Agency under Contract No. MDA903-75-C-0180.

REFERENCES

1. J.N. Holmes, "The Influence of Glottal Waveform on the Naturalness of Speech from a Parallel Formant Synthesizer," IEEE Trans. Audio and Electroacoust., pp. 298-305, June 1973.
2. A.V. Oppenheim and R.W. Schafer, <u>Digital Signal Processing</u>, Prentice-Hall Inc., 1975.
3. J. Makhoul, "Linear Prediction: A Tutorial Review," Proc. IEEE, pp. 561-580, April 1975.
4. O. Fujimura, "An Approximation to Voice Aperiodicity," IEEE Trans. Audio and Electroacoust., pp. 68-72, March 1968.
5. A.W.F. Huggins, R. Viswanathan and J. Makhoul, "Speech-quality testing of some variable-frame-rate (VFR) linear-predictive (LPC) vocoders," J. Acoust. Soc. Am., pp. 430-434, Aug. 1977.
6. F. Itakura and S. Saito, "Analysis Synthesis Telephony Based upon the Maximum Likelihood Method," Reports of 6th Int. Cong. Acoust., Tokyo, Japan, Paper C-5-5, pp. C17-20, 1968.
7. F. Itakura, personal communication.
8. O. Fujimura, "Speech Coding and the Excitation Signal," 1966 IEEE Int. Comm. Conf., Digest of Technical Papers, p. 49, 1966.
9. Y. Kato, K. Ochiai, O. Fujimura and S. Maeda, "A Vocoder Excitation with Dynamically Controlled Voicedness," 1967 Conf. Speech Comm. and Processing, Cambridge, Mass., pp. 288-291, 1967.
10. D. Coulter, Application of Simultaneous Voice/Unvoice Excitation in a Channel Vocoder, U.S. Patent No. 3903366, 1975.
11. D.H. Klatt, "Structure of a Phonological Rule Component for a Synthesis-by-Rule Program," IEEE Trans. Acoustics, Speech and Signal Processing, pp. 391-398, Oct. 1976.
12. H.W. Strube, "Synthesis Part of a 'Log Area Ratio' Vocoder in Analog Hardware," IEEE Trans. Acoustics, Speech and Signal Processing, pp. 387-391, Oct. 1977.
13. M. Sambur, A. Rosenberg, L. Rabiner and C. McGonegal, "On Reducing the Buzz in LPC Synthesis," 1977 IEEE Int. Conf. Acoustics, Speech and Signal Processing, Hartford, Conn., pp. 401-404, 1977.
14. B. Atal, personal communication.

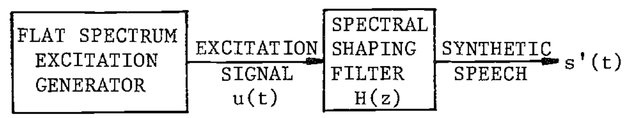

Fig. 1 Basic synthesis model.

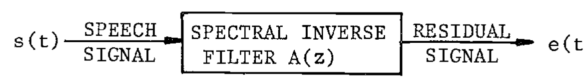

Fig. 2 Inverse filtering the speech signal to obtain a residual signal with a flat spectrum.

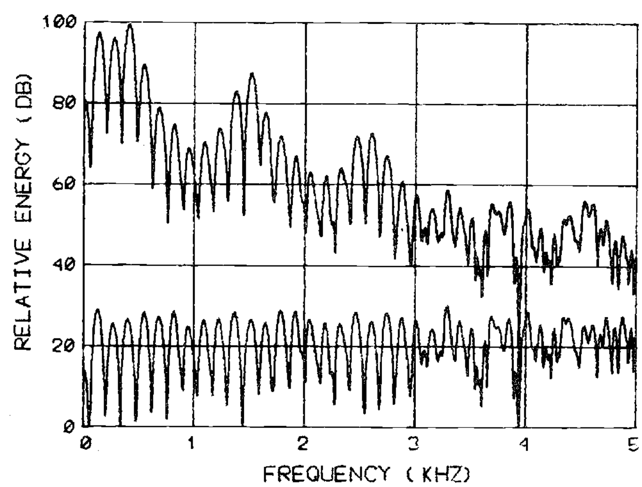

Fig. 3 Signal spectrum (top) and residual spectrum (bottom) for the vowel [I] in the word "list".

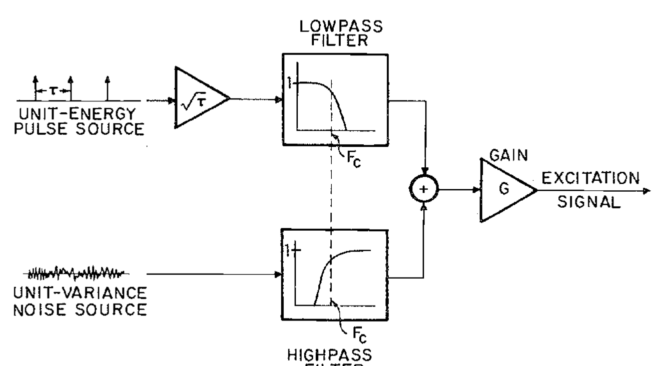

Fig. 4 Frequency-selective mixed-source excitation model.